NLP 감정분석 기반 마케팅 시장 분석

온라인 커뮤니티 특화 감성 사전 구축을 위한 새로운 용어 극성값 분석 시스템

빅데이터미네이터











주제: 온라인 커뮤니티 특화 감성 사전 구축을 위한 새로운 용어 극성값 분석 시스템

지난주주요피드백

- → 새로운용어가기존사전의 어떤용어와 비슷한지 찾는 기법 구현에 대한 고민 필요
- → 특허에 출원할 수 있는 아이디어
 - 1:1,1:n 매핑: 특정새로운단어가사전과 매핑될 수 있도록
 - 시간의 축: 단어의 뜻이 시간에 따라 변경 되는 부분 분석

구현 계획 변경 사항

- "네이트판"으로 커뮤니티 변경
 - → 크롤링을 진행하면 "too much results" 에러가 발생하는 FM코리아의 문제점
 - → FM코리아는 우회를 시도해도 우회 방지로 크롤링을 차단
 - → 네이트판은 10대 사용률이 높아 신조어가 가장 많이 출현해서 새로운 용어 추출에 용이함

오늘의 새글 3561개	최신 톡 톡 됐어!	요 많이본톡	추천 톡
[톡 채널]을 클릭하시면 해당 채널 글만 모아 볼 수 있습니다.			전체채널보기
제목	작성자	조회수	작성일
[19] 나요즘 정신상태가 너무 이상해 🛽	00	1	13:40
나 1학기 기말 40점인데 🖪	00	2	13:40
ㅎㅎ빛삭 🔼 🖪	00	4	13:40
[드루와] 마라탕 맛있어?? 🖪	00	4	13:40
배가 너무 아파 ㅠㅠㅠㅠㅠㅠㅠ 🖪	00	2	13:39
[댓글부탁해] 후드티 골라줄 사람 (2) 💌 🛛	00	9	13:39
[댓글부탁해] 키빼몸 112가 통통해보이긴 (1) ☑	00	4	13:39
나 유덩때 강남스타일나옴 🖪	00	4	13:38
등급컷 궁예해주라ㅠ (2) 🖪	00	8	13:38
진짜 초덩때로 돌아가면 잘 (1) 🖪	00	7	13:37
[댓글부탁해] 길냥이 드디어 데려옴 🖪	00	8	13:37
너네 공책정리 공부시간에 포함함? 🛛	00	4	13:37
고3들 술 먹어본 적 있어??? (2) 🖪	00	10	13:37

🗉 목록 〈 이전글 다음글〉
조회 42
▣ 목록 ⟨이전글 다음글⟩
조회 7

진행 사항: 새로운 용어 추출 및 분석

- KNU 감성 사전에 없는 단어 비교 추출
- 1. 크롤링 데이터 추출
- 2. KNU 감성 사전 데이터 추출
- 3. 두개를 비교하고 새로운 단어 추출해서 txt로 저장

'아기 냥 호빵 이 꽉 찬 개월 을 향 해가는 호빵 이 워낙 문동량 이 적어서 인지 많이 먹지 않아도 통통 냥 보는 사람 들 마다 이리 큰 냥 이 본적 없다 해서 걱정 입니다 키로 다이어트 시켜야 하나 친구 네 에서 좋아하는 자리 사랑 하 는 이불 동 굴욕 하 는 거 아님 목욕 하고 말리던 중탈출 옷장 안 목욕 하 면 드러나는 애기 애기 모습 매트리스 가로 지르기 보기 만 할 뿐 절대 탐내지 않음 잘 때 도 귀엽고 쩍 벌도 귀엽고제는 에 왕 콩깍지 뭘 해도 예쁜 우리 호빵 품 성한 한가위 보내세요항상 친절하고 진심 으로 대하 려 애쓰는데 나 는 늘 혼자 네 항상 친절하려 애쓰고잘 대해 주려 애쓰고 늘 진심 인데 항상 혼자 야 나를 피 한다는 것 도 느껴지고 불편하게 본다는것도 알겠는데 이유 를 모르겠어 뭘 노력해야 할까 내 가 연락 하지 않으면 나를 먼저 찾는이 하나 없고 광고 카톡 이외 엔 카톡 도 없어 다른 사람 들 은 카톡 에 전화 가 물리는데 내폰 은 늘 조용 씁쓸하다사진 은 어떻게 잘 찍 을 수 있는 거니 아마 이 게 살면서 가장 잘 찍었다고 생각 하는 사진 임나 도 사진잘 찍고시퍼 왜계속 찍어도 그대로 지 그리고 사진 에 사람 이 들어가면 배로 더 어려워짐 ㅠㅠ엄마 의 감정 쓰레기통 이었던 나 이런 마음 이 혼란 스러워 명절 을 앞두고 친정 에 가야 하나 사회 초년 생일 때 첫째 오빠 집 을 사야하는데 청약 이 없다며 내명의 로 된 청약 과 대출 을 빌리려던 친정엄마 초년 생일 시 작 한 후 부터 같이 사는데 생활비 집안일 요구 첫째 오빠 둘째 오빠 일 안 하니 용돈 보내라는 요구 결혼식 돈 다 가져가십 내 역도 모름 임신 당시 에는 무직 이었으나 생활비 요구 집안일 요구 출산 한지 한 달도 안된 차 매일 일 시키시고 자존감 갉아먹는 말 퍼

princ(wordinse)

'작다는', '컸던', '식으로회초리로', '가서보살피고', '학생회장', '신형', '힘내볼게요안녕하세요', '도둑취급했다', 각한거야', '김민주그리고', '합성이야', '토할뻔함하루', '안듣나', '비교되요', '엄마라고', '커버보고', 잖아요메이비식스움', '소식을', '여쭤보는거라', '무얼하든', '어머님은', '즐기다보니', '이것', '한공기', '들어왔어요대', '하지않냐면서요 실제로', '여자애들이라며', '낳기', '저희집에서', '해라', '지방까지주말마다', '장면인데', '프듀팬들', '아미들', '맟나봐', '소중했다그렇 게', '박아두어헤어진걸로', '컴백했어더블타이틀', '쨈이랑', '안된다', '오라함너무', '온갖욕', '긴글', '기관에' 에', '박지민', '대들', '싸우지않는데', '했는데요', '캄캄합니다저는', '컸고그오빠가', '가야한다는입장이고', '얼평하면서', '연휴', '먹으면', '먹어요그것', '돌봐주길', '보러가는', '간거랑', '의사자격없으면', '안키웠죠저도', '발언들도', '아웃오브안중폰', '찍어 놈경찰이', '가서같이', '욕하고', '일정이', '문제가', '세포들에서', '피곤해서', '선수', '있잖아', '우정도', '일등시민권을', '실베', '생 겼다', '입는', '자유인데암튼', '자매', '엄마랑', '네이트판', '이석훈과', '월클', '소집해서', '말안한거라고', '컴백곡', '좋아하던데' '열일루비가', '양해', '없긴했는데요', '슬픈자리에서', '무서우면', '만든', '넘어지자마자', '설치할수도', '봤는데', '잘해준다', '후회할거 라는둥', '니네', '하기전인데', '일주일에', '예정이니', '걔네', '몬스타엑스', '하게될수도', '쪽이', '아닌', '내라는', '했던', '마르진' '하는건가요당연히', '젖어있고', '어울리기도', '결혼은', '자기는', '힐', '이게', '모르쇠하네요', '우리도', '내주기엔', '밀어도', '소개해 주면', '원해야', '돈돈돈', '들어가겠습니다저는', '얘길', '길어지면서', '시간이였어요거의', '다니면', '틴트', '계시지만음', '개인의', '집게손가락만', '바꿈마지막은', '혼자계신', '눈빛이고', '모임메', '콧구멍', '잇어', '채팅메', '찿아보다가', '장사하시느라', '남자한테 '안그러는건가다', '잘해주시는', '그런거', '임신전', '거기다', '저에게도', '애교라며', '하나님이', '셔누랑', '데뷔부터', '고개돌려서' '상향', '시민권', '둘이', '작은엄마들', '탕이나', '반응은', '꿀팁이나', '새언니랑', '살아가고', '제니나', '계란후라이나', '포르투갈의 '세면대는', '게임을', '지랄났네발', '없어요시댁에', '남습니다', '재수없다고', '입장임작은', '걸림저', '라는게', '심리는', '왁자지껄한' '참여했어데자부는', '팬이', '합니다그리고', '자체로', '올라오면서항상', '식탐', '행복한일만'. , '동네에', '안가져다', '흙수저라', '찌면 '손이', '연로하심', '만지고', '싶을까요', '말하지말라그랬구요근데', '가져왔어진짜', '밥도', '좋아했고다투기라도', '않았어요엄마가', '여

진행 사항: 새로운 용어 추출 및 분석

- Soynlp 단어 추출
- 1. Soynlp 패키지를 사용해서 비지도 학습 기반의 명사 추출
- 2. 추출한 명사 중 새로운 단어 판별

```
from soynlp import DoublespaceLineCorpus
                                                                from soynlp.word import WordExtractor
                                                                word_extractor = WordExtractor()
# 문서 단위 말문치 생성
corpus = DoublespaceLineCorpus("natepann.txt")
                                                                word_extractor.train(corpus)
Ten(corpus) # 문서의 갯수
                                                                training was done, used memory 0.103 Gby 0.092 Gb
251
                                                                word_score = word_extractor.extract()
# 문장 단위 말문치 생성
corpus = DoublespaceLineCorpus("natepann.txt", iter_sent=True)
                                                                all cohesion probabilities was computed. # words = 1353
|Ten(corpus)| # 문장의 갯수|
                                                                all branching entropies was computed # words = 2561
                                                                all accessor variety was computed # words = 2561
451
```

진행 사항: 새로운 용어 추출 및 분석

- Kiwipy 단어 추출
- 1. 전처리 후 사전 미등록 명사 추출해주는 kiwipy 기능 사용
- 2. 새로운 용어를 명사 단위로 추출, 사용자 사전에 추가

- → 사용자 사전 추가 가능해서 사전에 등록 안된 명사들 추가 필요
- → 5일치 데이터라 미등록 명사 추출 양이 적었음, 데이터 양을 늘릴 필요
- → kiwipy 방식이 가장 정확성이 높았음

진행 사항: 새로운 용어 분석 아이디어

- Topic 분류 후 새로운 용어와 기존 단어를 매핑
- 1. 게시글을 보면 글마다 topic 분류가 되어있음
- 2. 이를 이용해서 데이터마다 topic을 분류해주고 레이블링하는 방식

```
독특 > 사는 얘기 > 모두드루와 시부모 -1, "결혼/시집/친정" 시부모 나1, "요리&레시피" 시부모 나1, "요리&레시피" 시부모 이, "이슈"
```

- → 단어마다 분류된 topic을 확인하고 분석에 사용하면, 글마다 단어가 다른 의미로 사용되는 문제를 해결할 수 있을 것으로 생각된다. Ex) 결혼/시집/친정 topic일때 '부정 ' 인 단어가 사는 얘기 topic에서 '긍정 ' 인 단어로 사용되는 경우
- → 새로운 단어가 들어오면 단어의 topic을 먼저 본 후, 단어의 연관 단어를 그 topic으로 분류된 단어만 추출해서 새로운 용어를 기존의 단어와 좀 더 정확하게 매핑시킬수 있도록 한다.

진행 사항: 새로운 용어 분석 아이디어

- 사용자의 특성에 따라서 달라지는 새로운 용어 의미 파악
 - 사용자의 나이, 성별, 커뮤니티 특성

- → 새로운 용어 입력시 미리 특성에 따라 분류된 커뮤니티에서 해당 단어를 검색하여 그 극성을 파악하고 어떻게 사용자에 따라 긍부정 극성이 달라지는지 수치화와 해당 단어가 사용되는 예시 문장을 제공
 - Ex) 멘토님 의견(요즘 사용되는 단어들)

문제가 된 단어 '오조오억'은 '아주 많다'는 뜻을 나타내는 신조어다. 그런데 최근 일부 남성 회원 중심의 온라인 커뮤니티(이하 남초 커뮤니티)를 중심으로 '오조오억'이 남성 혐오 단어란 주장이 제기되고 있다.

차주 계획

- 1. 크롤링 데이터를 늘려서 kiwipy 사용하여 새로운 용어 추출
- 2. 논문 작성 및 제출 한국멀티미디어학회

감사합니다:)

빅데이터미네이터

0/